DRTC Annual Seminar on Electronic Sources of Information
1-3 March 2000

Paper: CE

DIGITIZATION : VISION AND TASK

N. J. Deshpande, Librarian and Head, Dept. of Lib. & Inf. Science, Univ. of Pune, Ganeshkhind, Pune 411 007, & B. **M Pange**, Asst. Librarian, Univ. of Pune, Ganeshkhind, Pune 411 007

This paper discusses the problems and difficulties encounterd by traditional library systems while preserving and maintaining the library material. The paper tries to document in the form of DL, how to preserve and maintain the traditional and old records of library and information centres. It also discusses the various aspects of digital libraries like its definition, advantages and disadvantages of digitization. Digital library is a powerful tool of preservation, because there are tremendous amount of facilities, advantages and sustainability which in a far comprehensive and greater way exists over the traditional libraries. The paper in the main stream of thought compares and contrasts the traditional library and new form of digital library in order to arrive at a conclusion for a necessity and need for making digitization.

This paper discusses the need and advantages of Digitization. It highlights the process of digitization which will be helpful in undertaking the digitization projects and utilizing the funds in digitizing their resources.

1. INTRODUCTION

One of the significant aspects of science is the notion of change; that is change in observations, change in experimental set up, change in the methods, change in hypothesizing, change in theorization and the change in the overall perspective as well. Thus change is a ceaseless process; nothing remains stationary or static. In other words we can say that change is eternal and there is a flux. Digital library is a ceaseless change in the total librarianship, which is irreversible in its manifestation.

The original vision of the library as propounded by nineteenth century pioneers like Melvil Dewey and Charles A Cutter, was more than simply a set of pragmatic devices such as catalogue, classification system and reference desk procedures. It began in reality with a strong view of the cohesive and interrelated nature of knowledge itself, of humankind's accumulated social knowledge. To these pioneers, organizing information sources into a cohesive intellectual structure , regardless of the form of the structure, was

derivative of that preliminary vision of knowledge. Their efforts were shaped by what they assumed about that knowledge structure. When implemented in the form of bibliographic control practices that same structure provided a pathway to humankind's social knowledge.

There can be little arguments that the systems these people created (and which present day libraries still contend) have severe limitations with respect to modern information needs. Some of the limitations have arisen from their assumptions about knowledge itself, not only in how they viewed its organization (that liberally, in chiefly a two dimensional hierarchical structure, with monothetic classes) but also that there was only one way to organize it or that there was only one purpose for organizing. Other limitations arose from the technology they had at their disposal and inadequate ideas about the users habits of seeking information

Since there are some limitations in the traditional library system. librarians and information scientists started adopting new technology to provide better access to their collection. For several centuries, paper has been the primary medium for use in the conventional library system, because of its very attractive properties. New information handling techniques, storage and communication facilities have influenced the library system. One of the main reasons for the appearance of these new media is that they offer many types of facilities that paper based storage cannot afford. In his book on paperless publishing, Haynes offers a number of reasons for moving away from the use of paper towards the more extensive deployment of electronic media for the purpose of publication.

There are four basic steps for moving towards digitization.

1. The establishment of a paper based collection.
2. The extension of the basic system to accommodate the use of new publication media (new perspective)
3. The use of computer based methods to manage the system (management perspective)
4. Movement towards a totally electronic (digital) system.

2. WHY DIGITIZE ?

It is never too late to answer this question. The principal reason for digitizing materials is in collection is to overcome the twin tyrannies of time and space, the real barriers to the use of library collections. This is a paraphrase of Fred Kilgour' often repeated phrase " when and where the user wants it "meaning that the user should be able to choose the time when and place at which some orthodata is accessed. If our paper based collection are not digitized, this goal cannot be realized.

Another way to answer the question is to say that the objective is to reduce the time and effort the user must expend to gain access to orthodata. Libraries have paid little attention to efficiency from the user point of view.

The need for accessing the documents is increasing and this demand, in turn increases the risk of damage both gradual and catastrophic, to the documents. Although many types of damages can be repaired, the expenses involved can be considerable. In case of severe

damage, particularly with sound and moving image carriers the document is effectively destroyed. By digitizing a document the access needs of the majority of users can be met by using the digitized access copy. In short following are the reasons for digitizing the material :

a. To save space by replacing the printed originals
b. Provision of high resolution research tool with full text searching.
c. A quick overview or browsing tool, an aspect of preservation and conservation

2.1 Advantages of Digitization

Digital library is making it easier to write books, easier to save their content, and, in fact, easier to save everything being written. Another important aspect of digitization is preservation. Books, manuscripts, films, photographs and painting are decaying. Some of them are still useful, but the damage is beyond restoration. To preserve their information contents it is urgently needed to digitize them. Unesco sub-committee on preservation and access has identified the following advantages of digitization.

i. Copying - the potential to make copies of digitized information, using the same storage format or another digital format, without loss of quality.
ii. Automation - not only of the play-back of items requested by researchers, by using robotic storage systems, but of the process of making copies. As the document is represented by a string of binary numbers, it is possible to automate the process of copying and even to remove the need for human checking of the fidelity of the copy.
iii. Auto-Checking and Repairing - the possibility of automating the copying of digital information leads to the potential auto-checking of the condition of the collection and, if excessive errors are detected in a digital carrier, making a corrected copy of the material without human intervention.
iv. Searching - digitization offers the possibility of being able to search both local and remote catalogues and to create a web of links or pointers from the document accessed by the user to associated documents in the same or other collections. In addition, full text searches can be made. Similar search techniques are being devised for sound and image documents.
v. Access - the ability to send a digital signal over communications networks without loss of quality. The use of large scale robotic stores will enable the digitised collection to be accessible for 24 hours a day, with minimum staffing.
vi. Speed of Copying - within the digital domain, it is possible to accurately copy or transfer data at a very high speed. The provision of a hard-copy to a researcher can be quick. The future migration of collections to new carriers will be much faster than the initial movement to the digital domain.
vii. Quality - the potential to digitize an information carrier at a very high resolution or at a lower resolution as required. It is also very easy to generate a lower quality copy from the higher quality copy when required.
viii. Space Requirements - the density of storage of digitised information on the carrier can result in a major reduction in the shelf space required. This in turn reduces the space that requires archival climatic control with a consequent reduction in running costs.
ix. Future Migrations of the Collection - if a collection is copied to an analogue carrier, future migrations of the collection will cost (subject to inflation) the same as the

current migration. If the collection is copied to a digital carrier, future migrations can take advantage of the possibilities for automated transfers inherent in a digital format .

2.2 Disadvantages of Digitization

The major disadvantage perceived by many people is the cost. This can be summarized under four main headings:

i. Capital Costs - the equipment required to carry out the digitization process can be expensive to buy and often requires skilled operators if the best results are to be achieved.
ii. Storage Requirements - it is often considered that the carriers used to hold digitised information will require a very clean and climatically stable environment with a consequent increase in the energy consumption of the collection. There would also be a capital cost in the creation of any such controlled storage area.

iii. Running Costs - there are fears that a digitised collection may itself require frequent copying with the consequent labour, energy and new carrier costs. A safe life of only two to three years for digital information stored on magnetic tape and of three to five years for optical discs is feared.

iv. Preparation Costs - before capturing the document, it is necessary to prepare the material. In addition to any physical preparations that may be required, the preparation includes the ordering and indexing of the original material and the entry of the textual references into the data base by specialist staff. This can be an expensive process. The manpower requirement to copy the existing carriers to a new carrier is seen as being very high.

3. THE PROCESS OF DIGITIZATION

Digitization is the process of converting the microfilmed library material or printed library material into machine readable form using Optical Character Recognisition (OCR) / Intelligent Character Recognisition (ICR). There are basically three types of images which you might choose to scan ; photographic originals, which are made up of continuous tones of colour (which include greys for black and white photographs); half-toned images which are made up of a series of small dots, and often used in newspapers and magazines, and other illustration, usually known as line art, which are made up of blocks of colour usually black and white.

When an original is to be digitized directly, it is important to remember that the advantages of digital storage and processing must not be gained at the cost of reproduction quality, low durability, or lack of compatibility or future proofing of the information medium or of the hardware. A program specifying the technical and organizational steps involved in periodic migration, which can be constantly refined, should be part of the system design. The key elements in an image digitization system are :

Scanning
image enhancement

skew detection / correction
file format
compression
QC
Indexing
storage / retrieval
display
print
transmission

A document imaging system is an integrated configuration of hardware and or software components that produces pictorial copies (images) of office files, reports, publications and other source documents for storage, retrieval, dissemination and other purposes. This definition encompasses technologies and devices with widely varying characteristics and capabilities. The electronic images are typically produced by scanners that digitize source document for computer processing and storage. In most electronic image applications, the source documents are paper or paper like records like vellum or polyester sheets used for engineering drawings and maps. Like electronic document imaging systems, text storage and retrieval systems include computer hardware and software components which capture, process and store the textual contents of documents. Their significant advantages include very compact storage of textual information and the ability to retrieve documents by words or phrase that they contain. Following are the input steps in an electronic image installation

1.	Source document	
2.	Preparation	Pages made scanner ready
3.	Scanner	digitized images generated
4.	CPU Storage	temporary storage for digitized images
5.	Editing	text editing
6.	Magnetic disk	permanent storage of digitized images

The scanning process, which is properly termed document digitization divides a page into series of horizontal lines called scan lines. Each line is subdivided into small scannable units called picture elements, pixels, pels or in some cases dots. The pattern of scan lines and pixels employed by a given document scanner is termed a raster and resulting electronic images are often described as raster images. Using photosensitive components, a scanner measures the amount of light reflected by successively encounterd pixels and transmits a corresponding electrical signal to an image processing unit. Depending on equipment design, individuals scan lines and pixels are successively illuminated by a fluorescent lamp, an incandescent lamp, a light emitting diode or some other sources. Light reflected from the scanned pixels passes through a lens and onto a photo scanner which typically consists of a charge coupled device (CCD) array.

4. SPACE REQUIRED FOR DIGITAL IMAGE

Digitization offers a compact alternative to paper media. It occupies a considerable

amount of computer storage space. To store one page of information scanning requires more disk space as compared to character coded text. The storage requirement in electronic document images depends on several factors, including

- the linear dimensions of document
- the scanning resolution
- digitization mode employed
- compression algorithm used to use the amount of information

The following formula calculates the number of bytes required to store a single page of given size digitized at a specified scanning resolution

$$S = \frac{(H\ X\ R\ X\ B)\ X\ (W\ X\ R\ X\ B)}{8}\ X\ \frac{1}{C}$$

S = the storage requirement per page in bytes
H = the height of a typical subject document in inches or millimeters
W = the width of a typical subject document in inches or millimeters
R = the scanning resolution in pixels per inch or millimeters, along the documents horizontal and vertical dimensions
B = the number of bits utilized to encode each pixel and
C = an image compression factor

The calculation (H X R X B) X (W X R X B) yields the uncompressed page storage equipment in bits. The value of B is determined by the digitization of mode. With binary mode digitization of black and white documents, the most common scenario in electronic document imaging implementations, the value of B is one. Other possible values are eight for gray scale scanning of photographs, and twenty four for colour mode scanning. The most common resolution are 200 and 300 pixel per inch., but lower or higher values are possible

At 200 pixels per horizontal and vertical inch, each square inch of a page contains 40,000 pixels. At 300 pixels per horizontal and vertical inch the number of pixels per square inch increases to 90,000. The higher resolution image require 2.25 times more storage space.

Storage requirements are similarly affected by page size. Larger documents require additional pixel coding and occupy more storage space at all scanning resolutions.

Applying the above formula for a letter size page (11 X 8.5), we can calculate how much space will be required to scan at 200 pixels per horizontal and vertical inch. In this example the value of

H = 11
W = 8.5

R = 200
B = 1

S = (11 X 200 X 1) X (8.5 X 200 X 1)

We get the value of S as 3.74 million bits per page in uncompressed mode. When it is divided by eight we get 467,500 bytes. In this example, the uncompressed page storage requirement is 467,500 bytes. The amount is then multiplied by the reciprocal of the compression factor, that is a fraction with the anticipated compression as the denominator and the value one as the numerator

467,500 X 1/10 = 46,750 bytes

The image compression methodology employed by electronic document imaging systems are based on standard algorithm. The two widely encounterd examples are the CCITT Group III and CCITT Group IV algorithms developed for facsimile transmission by the Consultative Committee on International Telephony and Telegraphy. It yields a typical compression ration of 10:1 for office documents and 15.1 for engineering drawings.

5. CONCLUSION

Librarians play a critical role in giving priority for conversion into digitized form. They can develop strategies while selecting the material from the huge ocean of literature. The vision for the future calls for restructuring librarianship to enable readers in the vast resources of digital libraries.

While digitizing the material librarian should consider old and rare material, which needs more attention. They should also see whether the material is safely and successfully digitized and copyright permissions are obtained. It is also important to see how the digitized information can be retrieved when required. The decision to undertake the digital image capture in house or outside vendors has to be made. The process to bureau will depend upon the value and condition of the source material as well as the scanning equipments. Digitisation cuts across many traditional structures and involves. IT specialists, subject librarians, photographic and scanning technicians, service and special collections administrations. Future holds great promise for preserving the document in digitized form.

6. REFERENCES

1 Arms, William Y. Key concepts in the architecture of the digital library. D-Lib magazine, July 1995.

2 Arthur, Shiel and Broadhurst, Library material digitization demonstrator project British library : Library and Information Report-94, 1993.

3 Banks, Paul N. Preservation of library material In Encyclopaedia of library and information science. Ed. by Allen.

4 Browning, E.W. More training needed in book binding and conservation. Library journal v. 75 (Feb. 1 1950) pp. 190-91.

5 Chandel, A.S. and Devender Kumar. Preservation of reading materials in libraries in Lucknow Librarian v.13 no. 1, Jan- March 1981 pp. 1-12.

6 Digital full text library systems : a review of online projects and experiments at Georgetown University in Encyclopaedia of library and information Sc. Ed. by Allen Kent , pp.143-154.

7 Gagan, David and Martin Gillan, Scanning : a survival guide London: Information Partnership, 1993.

8 Gladey, Henry M. Digital library : gross structure and requirement : Report from a March 1994 workshop. Available at http://www.davison.net/catalog/books/

9 Graham, P.S. Building the digital library research library preservation and access at the heart of scholarship. Follett Lecture series, Leicester University, 19 March 1997. Available at : http://www.ukolin.ac.uk/services/papers/follet/ graham/paper.html

10 Goel, D.R. and Jaiswal, Kiran. Digital technology. In University news October 14, 1991, pp. 4-5.

11 Guidelines for digital preservation available at http://www.ahds.ac.uk/manage/framework.htm

12 Hampson, Andrew, Scanning in the right direction. Library technology 4(5) November 1999, pp.79-80.

13 Hedstrom M. and Montogomery S. Digital preservation needs and requirements in RLG member institutions : A study commissioned by the Research Library Group. California: Mountain View, Research Group, December 1998.
Available at : http://www.rlg.org/preserve/digpres.html

14 Horn, Andrew. Special material and services. Library trends v.4 no.2October 1955, pp.119-212.

15 Lundeen, Gerald. Conservation of library materials. Library trends v.30 no.2 Fall 1981, pp.175-317.

16 Saffody, William Electronic document : Imaging systems. London: Meckler, 1993, 183p.

17 Shoaf Eric C. Preservation and digitization : Trends and implications. Advances in librarianship. Ed. by Irene Godden v.20, 1996, pp.223-239 .

18 Smith, Martha M. Cooperative collection development for rare books among neighboring academic libraries. College and Research libraries. v.46, no.2, March 1985, pp. 160-168.

19 Tauber, Maurice. Conservation of library materials. Library trends v.4, no.3, January 1956, pp.215-334.

20 Williams, Lisa B. Selecting rare books for physical conservation : Guidelines for decision making. College and Research libraries v.46, no. 2, March 1985, pp. 153-159.